一种基于滑动窗口模型的 MOOCs 辍学率预测方法*

卢晓航 [1]　王胜清 [2]　黄俊杰 [1]　陈文广 [1]　闫增旺 [1]
[1](北京大学信息管理系　北京　100871)
[2](北京大学教师教学发展中心　北京　100871)

摘要:【目的】通过北京大学在 Coursera 平台上运行的课程数据, 对学生的辍学行为进行研究, 以期预测学生的辍学点和辍学行为, 改建教学慕课质量和方法。【方法】在课程数据基础上, 提取 19 个特征, 使用机器学习算法构建滑动窗口模型, 动态预测学习者辍学率。【结果】模型预测准确率高, 普遍在 90%以上, 效果稳定, 支持向量机(SVM)和长短记忆网络(LSTM)方法建模效果更好。【局限】课程数据选课人数偏多, 没有考虑其他课程数据稀疏问题, 模型的可移植性仍需要进一步考虑。【结论】使用滑动窗口模型建模, 能够帮助 MOOC 课程教师和设计者动态地追踪课程学习者辍学行为, 准确率高, 可以帮助教师通过快速的反馈来调整课程, 降低辍学率。

关键词: MOOC　辍学点　辍学率　滑动窗口模型　辍学预测
分类号: G434

1　引　言

MOOC 自 2011 年在美国兴起以来, 全球慕课的课程数量及用户数量都在逐年显著增长, 对高等教育生态、高校教学方法、管理制度及职业培训等领域的影响和冲击都在不断扩大, 国内外很多高校及机构都在不断进行慕课项目的尝试与实践。通常情况下, 每门 MOOC 都包括课程视频、课程论坛、课程维基、课程小测、作业和课程考试等模块。由于慕课是在线教学平台, 慕课平台能够把每门慕课的教与学的最原始数据都进行记录和保存, 为以数据为基础的学习行为分析研究提供极大的便利, 因而吸引了全球范围内的众多研究者开展慕课相关的数据分析研究。同时, 由于慕课是新兴的且在不断发展中, 慕课涉及的相关研究问题众多, 基于慕课的学习者学习行为分析是目前该领域的重要研究方向之一。

MOOC 具备开放、免费等众多优势, 由于不具备师生面对面交流等特点, 造成许多传统教学所没有的问题, 其中最引人注目的便是 MOOC 极高的辍学率。根据相关研究数据显示, 国内外大多数的 MOOC 课程结课率不足13%[1]。如何提高 MOOC 的课程通过率, 对辍学行为进行预测并进行干预, 分析辍学原因, 改善课程质量和在线教学方法是 MOOC 教师和设计者十分关注的问题。

本文从北京大学的 2013 年秋季、2014 年春季及 2014 年秋季在 Coursera 平台上开设的三个学期多门慕课课程数据作为数据样本集。针对学习者辍学问题, 统计分析辍学时间和开始时间的特点, 并提出一种滑动窗口模型, 动态地预测课程学生的整体辍学率, 帮助教师提升课程质量, 及时与潜在慕课退学者进行沟通, 提供帮助和反馈, 进而提升课程的结课率。

通讯作者: 王胜清, ORCID: 0000-0002-7164-073X, E-mail: wangsq@pku.edu.cn。

* 本文系教育部在线教育研究中心教育基金(全通教育)重点项目“慕课在线教学组织方法实证研究”(项目编号: 2016ZD301)的研究成果之一。

2 MOOC 辍学率问题的相关研究

对于慕课学习者什么时候会离开课程、辍学这一问题，国内外已有许多学者在研究，现有研究的分析数据可分为两大类：论坛数据分析和点击流数据分析。本文挑选几个典型的辍学率研究进行分析，总体而言，对于辍学率问题的研究虽然比较多，但并没有出现标准化的学界共同认可的研究方法，多数研究处于探索尝试中，使用不同的模型和方法来提高慕课辍学率的预测准确率。

Amnueypornsakul 等[1]使用学习者的点击流数据预测学生是否会辍学。将每位学习者的每周学习行为形成序列，如<wwaaws>表示学习者在课程中的完整行为：浏览课程 wiki、浏览课程 wiki、做测试、做测试、浏览课程 wiki、提交测试。将每位学习者的每周学习行为形成序列，定义三种学习者：活跃、弃学(即没有学习行为)、不活跃(学习行为序列元素少于 2 个)。在定义“辍学”时，分为三种情况：将不活跃的学习者归为辍学者，将不活跃的学习者归为活跃学习者，将不活跃的学习者以 0.5 的概率归为辍学者、0.5 的概率归为活跃学习者。利用 SVM 进行模型构建时，也分为两种情况：剔除不活跃用户进行预测模型构建；包含不活跃用户进行预测模型构建。至此，共构建 6 种模型。每个模型基线(Baseline)为学习者本周学习行为序列小于等于 1 时预测下周辍学情况。结果显示，剔除不活跃的用户进行预测模型构建时，准确率有很大提升；如果包含不活跃的用户进行预测，将不活跃的学习者定义为不属于辍学时准确率较高，但依然比基线低。

Sinha 等[2]利用视频点击和论坛数据，构建学习者的活动序列，寻找能够代表学生积极或消极参与课程的足迹序列。首先构建学习者每周的活动序列，从中提取 n-gram 序列、视频观看活动序列和论坛交互序列，以此探究何种序列能预测学习者辍学以及何种序列能让学习者保持热情参与 MOOC 学习。并从两个方面进行实验：一周之内的行为如何影响下一周的辍学；自第一周的累积学习行为如何影响下一周的辍学。此外，运用社会网络分析方法分析学习者行为序列图。结果显示，辍学的学生有更少的节点、边、强连通分量和自回路，辍学行为更受最近几周学习行为的影响。且大多数辍学的学生是在课程开始的几周后开始上课，有更稀疏的活动图。有两种可能的解释：这些辍学的学生有特定的信息需求，获取所需信息后不再上课；后来加入的学生由于之前的材料和作业太多，难以跟上课程而放弃。

Taylor 等[3]预测辍学时使用不同的机器学习方法做了很多尝试，包括逻辑回归、支持向量机、深层信念网络、决策树、隐马尔科夫模型等。将辍学定义为学习者不再提交任何作业和测试，并筛选学习者的 14 周学习行为数据进行训练和测试。将学习者分为 4 类：消极参与、参与编辑 wiki、参与编辑论坛、积极参与(既编辑 wiki 又编辑论坛)，并分别对这 4 类学习者进行建模。研究者提出超前滞后(Lead and Lag)的预测模式，即给定一周 i，使用前 i 周数据预测剩下的 14–i 周。在预测中，探讨各个特征的权重，以及哪些特征能在学习的开始阶段预测其能否坚持学习到达课程结束。结果显示，对于消极参与群体的预测准确率最高，而由于数据量不足，对编辑 wiki 和积极参与的群体的预测准确率较低。但是，编辑 wiki 这一特征能够较好反映学习者能否坚持学习到课程结束这一行为。如数据量充足，各种构建模型的方法的预测准确率差别不大。并且，对于某周预测，最近 4 周的数据更有预测性。在预测特征方面，结果显示，对于那些熟悉 MOOC 的人能够提供有较好预测性的特征，与提交作业、测试相关的特征预测性都很高，发帖长度则比发帖数更有预测性，以及与合作社交相关的特征如 wiki 和论坛等在预测中十分重要。

此外，一些相关研究也有一定启发意义。如 Kloft 等[4]使用点击流数据和机器学习算法预测辍学行为，并且在预测过程中，对每个特征向量做辍学预测性的检验；结果显示前 8 周预测效果不好，后面每周预测效果上升；分析原因为前几周数据量少，并建议加入论坛数据进行预测。Sharkey 等[5]则详细描述使用机器学习技术预测辍学的迭代过程，并通过研究得出带有预测性的特征以及它们的相对权重。结果显示，机器学习模型在预测辍学方面高于平均水平，预测因素也与人们的期待相同，都是能够显示学生是否热情参与的变量。Yang 等[6]的研究显示社交因素(论坛)对辍学确实有影响，并给予 MOOC 设计者启示。

通过上述研究，发现学生的点击流数据分析是目前辍学率问题研究的主要方向，基于点击流数据构建更为优化的数据分析模型是本文的重点研究方向。

3 建模与实验

选取北京大学 2013 秋季、2014 年春季及 2014 年秋季开设的 3 个学期、共计 5 门的 MOOC 课程的日志数据作为分析对象。经过研究分析发现，MOOC 课程中主要有两种模式：一门课重复多个学期开设，每个学期的课程内容相同；一门课程分成上下两个学期开设，前一学期课程是后一学期课程的基础。据此作为挑选课程的原则和依据，基于这两种模式，选择已开设 3 个学期的生物信息学和已开设两个学期的社会调查与研究方法(上)、(下)共 5 门课程。一方面因为这 5 门课程满足这两种课程模式，另一方面这 5 门课程的学习者很多，学习行为的数据量大，便于开展研究。

3.1 数据来源

为能够对 MOOC 辍学情况有更为直观的认识，本文统计了这 5 门 MOOC 课程的注册人数、记录有成绩、成绩大于 0 的人数、最终成绩大于 60 及每门课程的通过比例，如表 1 所示。

表 1 课程注册人数与通过比例

课程 ID	课程名	注册人数	记录有成绩	最终成绩大于 0	最终成绩大于 60	通过比例(%)
methodologysocial2-001	社会调查与研究方法(下)	3 566	3 184	371	185	5.1879
methodologysocial-001	社会调查与研究方法(上)	7 836	6 051	6 051	255	3.2542
pkubioinfo-002	2014 生物信息学 002	16 714	15 790	1 268	510	3.0513
pkubioinfo-001	2013 生物信息学 001	18 367	18 367	1 620	520	2.8312
pkubioinfo-003	生物信息学–导论与方法	16 958	16 072	909	360	2.1229

表 1 显示，这 5 门课程的通过比例均不高于 6%，与其他相关研究中所得情况基本一致。这无疑是 MOOC 辍学率高、通过人数占比小的一个局部反映，某种程度上来说，也更凸显了对 MOOC 辍学进行预测的必要性。

3.2 特征提取

每门 MOOC 课程都包括教学视频、小测、论坛等不同的学习模块。为了能够更准确地对学习者的学习行为进行预测，本文抽取多个学习行为数据，包括视频点击流、课堂测验和课程论坛。

在获取具体数据时，通过对日志文件中的 URL 提取关键字，辨别学习者利用的学习模块，获取其一段时间内的学习行为数据。

学习者将会通过观看视频、参与论坛、进行测试等多种形式参与到课程的学习中。参与过程具备以下两个重要的特征。

(1) 课程进度是以周为单位推进的，学习者可以在一周内的任意时间段完成该周的学习任务。事实上，可以在更小的时间单位下，讨论对于学习行为的追踪。

(2) 学习者的学习行为是单向为主，偶有双向行为的模式。学习行为中大部分是单向地接受知识的过程，反映在点击、浏览等行为特征上；而双向即学习者交互活动的部分是相对小众的，占比较小。这也是许多当前的课程引入同伴评价作业形式，以加强双向参与的原因。

在获取点击流数据时，发现学生的在线学习行为主要聚焦于观看视频、查阅资料和完成作业相关的行为。据此提取观看视频、查阅资料的内容获取的 8 个学习行为特征及在线完成作业的相关的 3 个行为特征。

关于是否将课程论坛数据引入为特征数据，研究者之间的观点存在分歧。Amnueypornsakul 等[1]认为，只有 5%-10%的学生会参与论坛，大多数学习者没有任何的论坛行为数据。对这些不参与论坛的学习者，利用论坛数据进行预测并不恰当，因而决定不使用论坛数据，只使用点击流数据。基于此，本文对论坛参与行为与学习者是否取得成绩进行相关性检验。相关性检验公式如下所示。

$$R = \frac{(SET1 \cap SET2)}{SET2} \quad (1)$$

本文定义至少参与 1 次论坛行为的用户集合为 *SET1*，而取得成绩的用户集合为 *SET2*，发现每一门课程中，*SET1* 与 *SET2* 都具有很高的重合率，如表 2 所示。

表 2　有学习成绩与有论坛行为重合率统计

课程 ID	有论坛行为	成绩大于 60	有论坛行为且成绩大于 60	有论坛行为在有成绩学习者中占比(%)	有论坛行为在成绩大于 60 的学习者中的占比(%)
pkubioinfo-001	2 645	580	511	68.3333	88.1034
pkubioinfo-002	1 425	508	395	54.5741	77.7559
pkubioinfo-003	1 523	358	316	66.9967	88.2682
methodologysocial-001	1 165	290	269	17.8318	92.7586
methodologysocial2-001	326	203	153	64.4205	75.3695

以 methodologysocial-001 课程为例, *SET1* 与 *SET2* 的重合率约为 92.8%, 也就是说, 该课程中最后通过课程的学习者中有约 92.8%都参与了该课程的论坛。

因而, 本文认为, 论坛参与行为对于学习者是否会坚持学习有明显的预示–标识作用, 所以将学生参与论坛的数据加入到预测模型中。

3.3　特征列表

经过研究分析, 本文提取的特征数据项主要有 19 项, 如表 3 所示。

表 3　提取特征列表

特征	字段	数据类型	备注
点击流	page_view	Int	查看网页
	page_view_quiz	Int	查看测试页面
	page_view_forum	Int	查看论坛页面
	page_view_lecture	Int	查看视频页面
	page_view_wiki	Int	观看课程 wiki
	viedo_view_times	Int	观看视频次数
	video_pause_times	Int	视频暂停次数
	video_pause_speed	Float	播放速率
作业测试	try_hw	Int	尝试作业次数
	try_quiz	Int	尝试小测次数
	try_lec	Int	尝试讲座次数
论坛行为	view_forum	Int	查看论坛
	thread_forum	Int	查看线程
	post_thread	Int	创建线程
	post_comments	Int	发表评论
	Upvote	Int	点赞
	Downvote	Int	反对
	add_tag	Int	增加标签
	del_tag	Int	删除标签

3.4　学习周期

学习周期是指学习者的学习开始与结束时间, 为了能够统一标准, 需要对学习开始时间和结束时间进行统一定义, 并以周为单位进行划分。

由于课程数量众多, 起始时间不一, 本文将每个学习者的开始时间定义为第一条视频点击数据出现的时间。为确定学习结束的时间, 且使得最后几周仍有足够的数据用以预测, 以维持较好的预测准确率, 对这 5 门课程中取得成绩的学习者最后一次学习行为发生的时间进行统计, 并选择取得成绩的人中前 80%的人结束学习的时间作为课程的结束时间。

如图 1 所示, 2013 年社会调查与研究方法(上)课程中取得成绩的人数中, 学习行为结束的时间中最早的为第 9 周, 最晚的为第 19 周, 其中有 80%的学习者在第 14 周便结束了学习, 因此将第 14 周作为 2013 年社会调查与研究方法(上)课程的结束时间, 使得课程末期的预测准确率得到保障。其他课程与此类似。

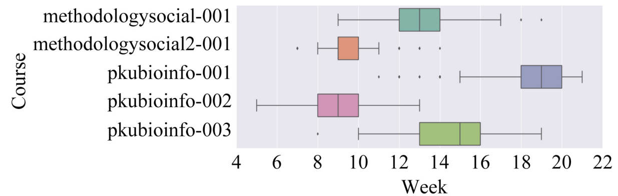


图 1　课程开始结束时间示意

3.5　在线学习人数

在线学习人数是指每周有学习行为发生的人数, 以这 5 门课为例进行统计, 结果如图 2 所示。这 5 门课程的学习人数均呈现在初期迅速上升, 之后从急剧下降转变为平缓下降的态势。此外, 这些课程最后阶段仍在学习的人数远远低于学习人数最多的初期, 在一定程度上也反映出 MOOC 辍学现象的普遍性。

3.6　开始点、辍学点及辍学

对每个人的开始时间和辍学时间进行更深入的统计分析。由于 MOOC 在课程期间任何时候都可以开始学习和结束学习, 因此每个人在 MOOC 上停留的周数

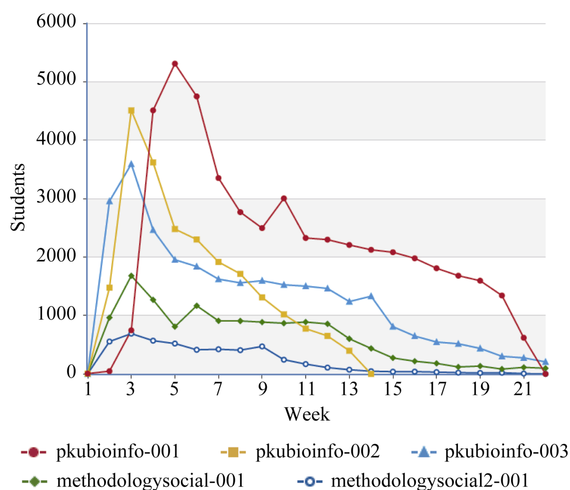


图 2　课程在线学习人数变动示意图

差距很大。本文将每个学习者某周是否有学习行为定义为 1 或者 0(有学习行为定义为 1)，这样可以获得每个学习者每周是否出现的特征序列。如果某个学习者在第三周进入课程，则其特征序列为 0-0-1-……。定义第一次出现 1 的周为开始点；如果某个学习者在某周之后不再出现，即序列为……1-0-0……-0，则定义该周的下一周为辍学点。同时笔者认为从辍学点开始，这个学生已经从本门课程辍学，即从辍学点开始学习者不再出现。研究着重观察每个辍学点的辍学人数以及开始点与辍学点之间的关系。

通过观察这 5 门课程的每周辍学人数，发现在课程开始两周后，课程的辍学人数会达到一个高峰，在课程临近结束时，辍学人数又会有一定的上升，课程中期的辍学人数则比较平稳，如图 3 所示。

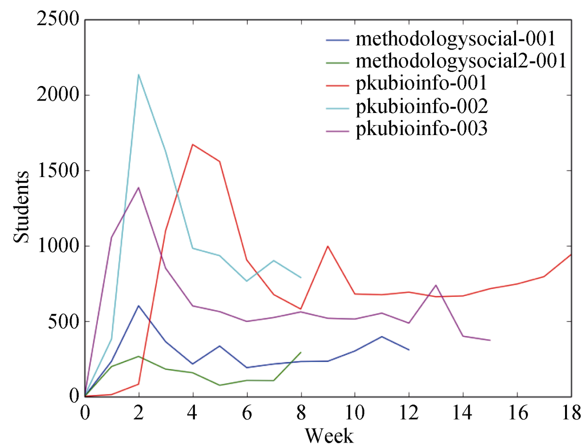


图 3　课程辍学人数变动示意图

原因可能为，学习者在学习该课几周后，对该课有了一定的了解，会根据该课是否适合自己而做出是否辍学的选择；在课程结束时，并不在意期末考试或能否取得成绩的人可能做出辍学的选择，也会有一部分学习者因自认为无法获得证书而辍学。中间阶段辍学人数的平稳则恰恰可能反映了正常的学习者流入和流出，也反映出坚持到中间阶段的学习者辍学概率较小。

图 4 考察开始点与辍学点之间的联系。图 4 中纵坐标表示每周辍学率(即辍学人数占某周开始的总人

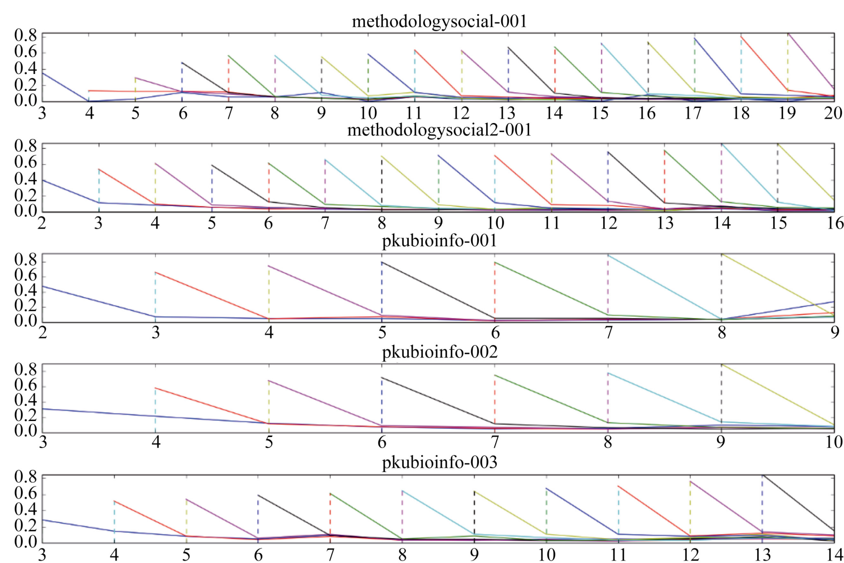


图 4　开始点及辍学点关系

数的比率)，横坐标表示时间，即当前是第几周，从左到右的折线表示第 n 周开始。通过比对图表之中的共性，发现：开始的时间越晚，学习一周就辍学的比率越高；开始的时间越早，坚持到后面几周的比率越高。说明开始的时间越早，越有可能坚持到最后。开始的时间越晚，在下一周辍学的可能性越大。此处可援引 Sinha 等[2]的解释，一种可能是开始较晚的人由于之前的材料和课程内容太多而难以跟上，另一种可能是开始较晚的人更可能是专为谋求某种特定的信息而来，获取后即不再学习。

3.7 滑动窗口模型的构建

在上述分析定义的基础上，重点讨论滑动窗口模型的构建，用以预测学习者是否辍学。该模型将整个学习周期视为一个连续序列，通过之前若干周的特征向量，预测未来几周学习者是否会参与课程。如图 5 所示，滑动窗口模型分为前后两个窗口，第一个窗口长度为 w_1，即当前周的前 w_1 周，如果当前周为第 n 周，则窗口内为第 n–w_1 至 n–1 周。第二个窗口长度为 w_2，即包括当前周及之后的 w_2 周，如果当前周为第 n 周，则窗口内为第 n 周至第 n+w_2–1 周。该模型使用窗口长度为 w_1 周内的 19 维特征向量的点击流数据，预测之后 w_2 的标签，以 w_2 的标签(Label)为分类目标，即辍学或者不辍学；通过窗口的滑动，对课程的每一周进行学习者是否辍学的预测。因此辍学是指 w_2 长度周学习者是否有学习行为，没有即为辍学。由于该模型只关注当前窗口内学习者的特征向量，而没有将学习者从前至后的所有学习行为联系起来，因此并不重点关注个人学习者的辍学和什么时候辍学，只关注学生有无学习行为的总体情况，即当前窗口内，会有多少人辍学。

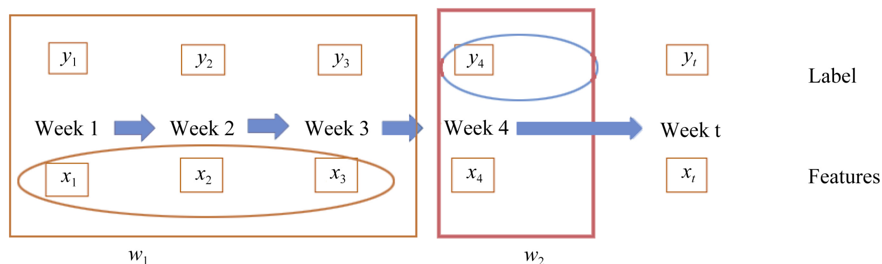


图 5 滑动窗口示意

由于 MOOC 课程中大部分的学习者都不会坚持到最后，因此定义基线为预测所有学习者下周不会再出现在课程中，进而比较各个模型相对基线的改进。基线预测会导致很多的错误，但也是最简单直接的预测。这一预测方式在现实中也有大量的应用，如教师会伴随着课程的进行不断地广发邮件鼓励学习者继续学习，即使该学习者在下一周不会流失。引入机器学习的方法可以改进这种策略。对于预测模型，采用逻辑回归(LR)[7]、支持向量机(SVM)[8]、多层感知器(MLP)[9]、长短期记忆(LSTM)[10]作为分类器。

在前述 5 门课程上进行滑动窗口模型的实验。选取 w_1 的值时，如果 w_1 太短，如 w_1=1，即前一周预测后一周，模型粗略简单；如果 w_1 太长，又过于冗余，根据以往研究，选择 w_1=3, w_2=1，并对模型进行 5-fold 交叉检验。把数据分成 5 个部分，选择 1 个作为测试数据，剩下 4 个作为训练数据，实验重复 5 次，平均实验结果如图 6 所示。通过分析，可见以下特点。

(1) 基线准确率总体偏高。这是因为每周都会有较多的人辍学，普遍情况下辍学率在 70%，因此基线的准确率就会偏高。

(2) 课程开始和结束时期的辍学率处于峰值。早期离开的人多，因为学习者在早期会尝试学习，若有不适合自己等原因就会放弃学习；而临近课程的末期，也会有很多人因为课程的压力而放弃，在此时进行适当干预和鼓励或许可以显著降低辍学率。

(3) 普遍情况下，机器学习方法预测效果好。在不同的机器学习方法中，逻辑回归代表基本情况下机器学习模型的预测能力。相比简单地认为学习者不能坚持学习，机器学习能够更好地识别出能坚持学习的学习者。但其预测能力有限，需要更多的特征数据才能达到更好的效果。

(4) LSTM 和 SVM 效果较好。相比多层感知器和

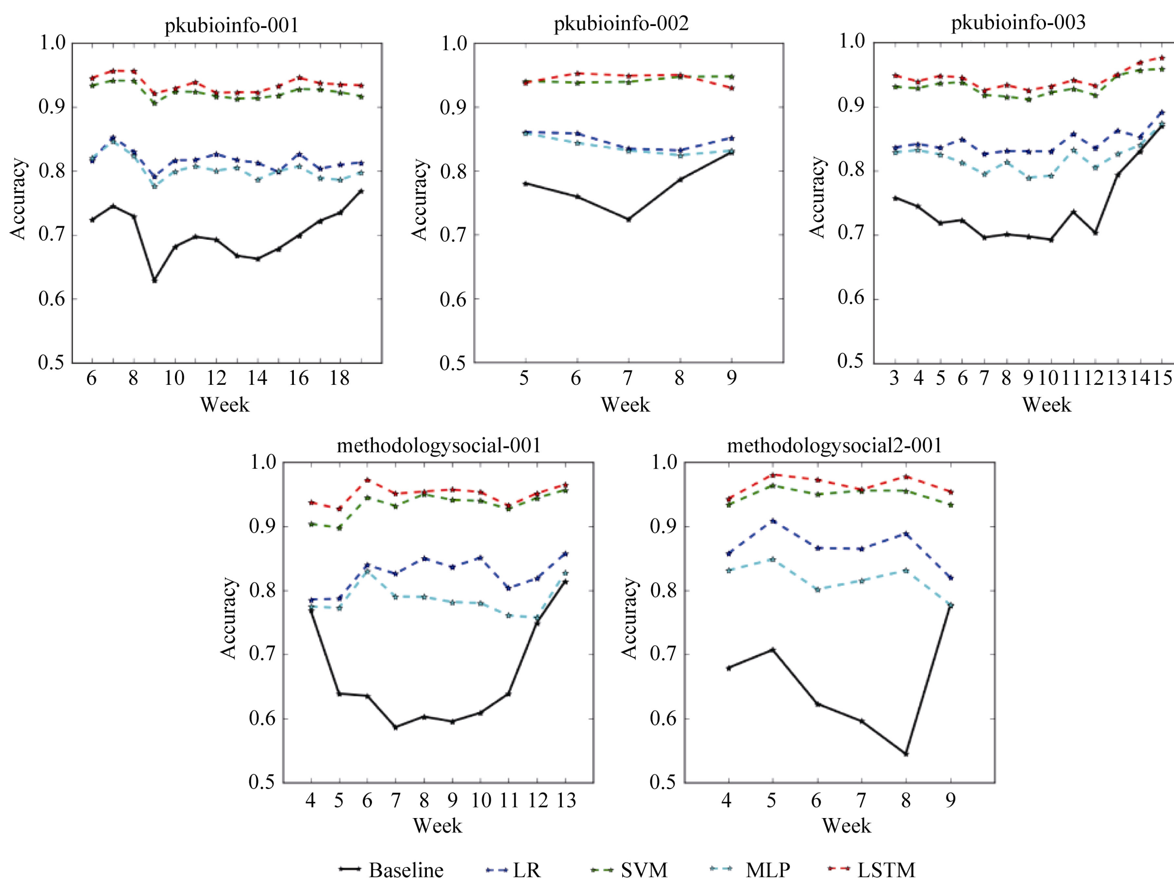


图 6　各个模型在 5 门课程上的预测准确率(w_1=3, w_2=1)

逻辑回归，这两种方法预测能力更好，效果稳定，不受数据量的影响。

进一步将后置的窗口扩大，即预测接下来几周的表现，如图 7 所示。取 w_2=3，这时会出现 000,001,010, 100,011,101,110,111,共 8 种情况，对应不同的分类 0-7。对后置窗口的情况，本文使用 Baseline1 代表预测接下来 3 周不会有学习行为发生的情况，即 000; Baseline2 则代表预测连续3周都有学习行为发生的情况，即111。

经过实验，发现该结果呈现以下的特点：学习者未来三周的学习行为中，000 比例相对较高，111 的比例相对较低，并且课程开始和课程结束时 111 的比例都是最低的。即在课程开始和结束阶段，进行连续周学习的学习者比例很低。此外，在预测后面多周学生是否辍学情况下，SVM 将其简化为多分类问题是有效的，因为 000 的比例相对较高。而 LSTM 和 SVM 虽然方法上存在差异，但是结果均保持较高的准确率。

滑动窗口模型初步解决了对于课程不同阶段的监控、预测问题，并发现 MOOC 课程的开始阶段和结束阶段是对于学习者最具挑战的阶段，容易造成辍学。在这一阶段辍学的原因与 MOOC 本身特点有关，学习者在早期可能只是了解课程，在发现不适合自己之后，会选择离开，这是前期辍学率较高的原因。因此课程本身如果能够激发学习者的兴趣，并且在前期保持对学习者的关怀和帮助，将会有利于课程进入相对平稳的时期；而在学期结束时，由于期末考试，会有不少的学习者选择离开课程。因此如果能够在学期末采取一些能够鼓励学习者完成最后考核的策略，例如复习课程等，则会有利于在学期结束时降低辍学率，鼓励他们最后取得成绩。另外，虽然本研究仅选取 5 门课程的数据，但实验时课程名是变量，可以修改为任意课程，因此把本模型应用于其他 MOOC 课程进行辍学率预测，如大学化学、计算概论、刑法学、人群与网络课程中，发现依然是有效的。

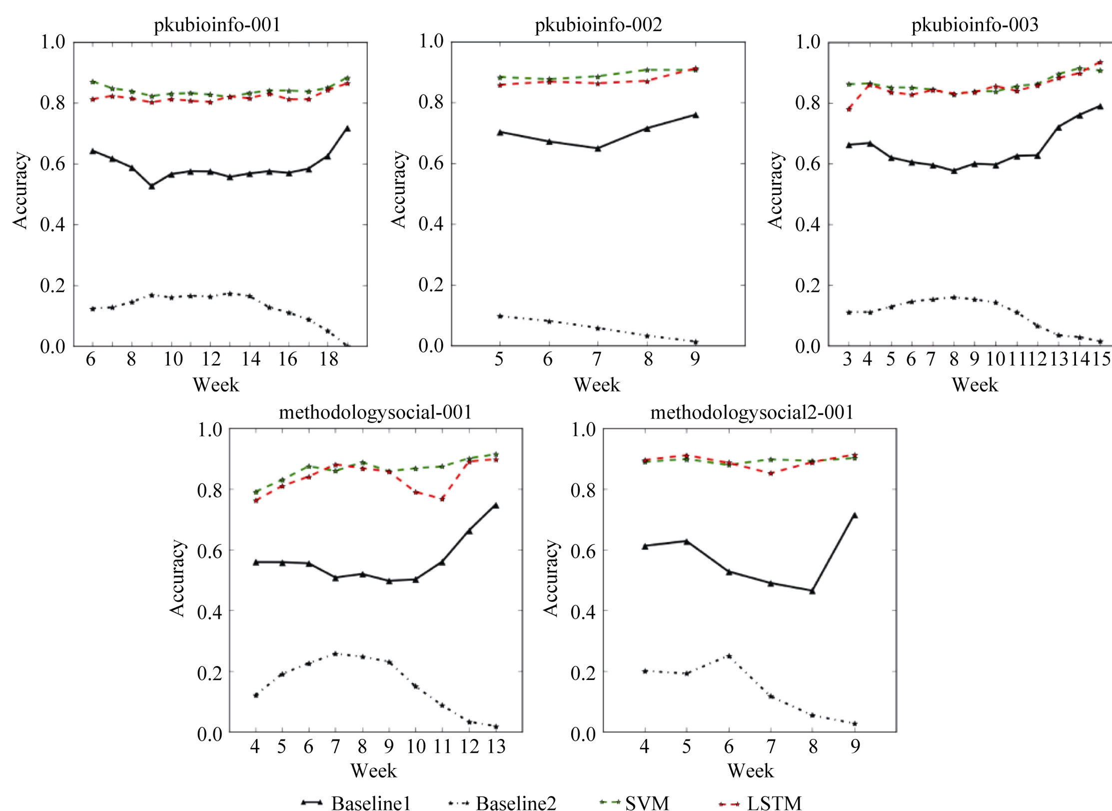


图 7　各个模型在 5 门课程上的预测准确率(w_1=3, w_2=3)

4　结　语

本文从北京大学MOOC课程数据入手，对其进行清洗，根据以往研究以及数据特点，抽取19个特征对课程数据进行研究。在分析课程的在线学习人数变化，学习者开始点和辍学点的特点等数据的基础上，通过构建滑动窗口模型，能够有效预测一门MOOC课程的整体学生辍学情况，发现MOOC课程辍学情况严重，且开始时间越晚的学习者辍学的概率更高。结果表明，机器学习方法及滑动窗口模型对预测学生的辍学有很高的准确率，能够帮助教师追踪课程，预测辍学学生，把握课堂进度。该模型能在课程进行中帮助设计者和教师通过快速反馈调整课程，以期降低辍学率。

未来研究重点将放在不同课程间辍学模式的异同，观察分析不同课程间学习者辍学动机的异同，以及预测学习者能否取得成绩。并进一步在更多的数据上进行实验，完善模型，给予不同类课程以不同的反馈建议，辅助MOOC课程教师更好地设计自己的课程，帮助教师与学生更有效及时沟通，对潜在辍学的学生给予干预，帮助学习者更有效地学习，最终降低MOOC的辍学率，提升在线教学质量和教学效果。

参考文献：

[1] Amnueypornsakul B, Bhat S, Chinprutthiwong P. Predicting Attrition Along the Way: The UIUC Model [C]// Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Doha, Qatar. Association for Computational Linguistics, 2014: 55-59.

[2] Sinha T, Jermann P, Li N, et al. Your Click Decides Your Fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions[OL]. arXiv Preprint. arXiv:1407.7131, 2014.

[3] Taylor C, Veeramachaneni K, O'Reilly U M. Likely to Stop? Predicting Stopout in Massive Open Online Courses[OL]. arXiv Preprint. arXiv: 1408.3382, 2014.

[4] Kloft M, Stiehler F, Zheng Z, et al. Predicting MOOC Dropout over Weeks Using Machine Learning Methods[C]// Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, Doha, Qatar. Association for Computational Linguistics, 2014.

[5] Sharkey M, Sanders R. A Process for Predicting MOOC Attrition[C]//Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs,Doha, Qatar. Association for Computational Linguistics, 2014: 50-54.

[6] Yang D, Sinha T, Adamson D, et al. "Turn on, Tune in, Drop out": Anticipating Student Dropouts in Massive Open Online Courses[C]//Proceedings of the 2013 NIPS Data-driven Education Workshop. 2013: 11-14.

[7] Lipsitz S R. Categorical Data Analysis[J]. Statistics in Medicine, 1992, 13(11): 1791-1792.

[8] Cortes C, Vapnik V. Support-Vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.

[9] Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms[J]. American Journal of Psychology, 1962, 7(3): 218-219.

[10] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1-32.

作者贡献声明:

卢晓航: 部分数据统计, 论文部分文字撰写及统稿;
王胜清: 研究工作总体设计, 论文终稿审核;
黄俊杰: 数据处理建模及实验;
陈文广: 数据分析方法设计及优化;
闫增旺: 部分数据统计及论文部分文字撰写。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 1300016603@pku.edu.cn。

[1] 卢晓航. methodologysocial-001.zip. methodologysocial-001 数据.

[2] 卢晓航. methodologysocial2-001.zip. methodologysocial2-001 数据.

[3] 卢晓航. pkubioinfo-001.zip. pkubioinfo-001 数据.

[4] 卢晓航. pkubioinfo-002.zip. pkubioinfo-002 数据.

[5] 卢晓航. pkubioinfo-003.zip. pkubioinfo-003 数据.

[6] 卢晓航. pkuic-001.zip. pkuic001 数据.

[7] 卢晓航. peopleandnetworks-001.zip. peopleandnetworks-001 数据.

[8] 卢晓航. chemistry-002.zip. chemistry-002 数据.

[9] 卢晓航. criminal law-001.zip. criminal law-001 数据.

收稿日期: 2017-02-27
收修改稿日期: 2017-04-06

Predicting Dropout Rates of MOOCs with Sliding Window Model

Lu Xiaohang[1]　Wang Shengqing[2]　Huang Junjie[1]　Chen Wenguang[1]　Yan Zengwang[1]
[1](Department of Information Management, Peking University, Beijing 100871, China)
[2](Center of Faculty Development, Peking University, Beijing 100871, China)

Abstract: [**Objective**] This paper aims to improve the MOOCs curriculum quality and pedagogy by analyzing the dropout behaviors with data from the MOOC of Peking University on Coursera. [**Methods**] We extracted 19 major features from the logs and then constructed a siding window model to predict the dropout rates. [**Results**] The precision of the proposed model was maintained above 90%. The SVM and LSTM methods further improved the performance of the proposed model. [**Limitations**] The new method needs to be examined with smaller sized courses. [**Conclusions**] Predicting dropout rates could help us improve the course quality effectively.

Keywords: MOOC　Dropout Point　Dropout Rates　Sliding Window Model　Dropout Prediction